Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Complexes

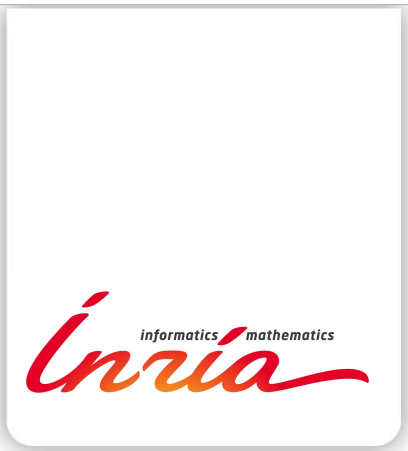
Tom Dreyfus, Valérie Doye, Frédéric Cazals

Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Complexes

T. Dreyfus and V. Doye and F. Cazals

RESEARCH REPORT
N° 8118
October 2012
Project-Team ABS

ISSN 0249-6399 ISRN INRIA/RR--8118--FR+ENG



Probing a Continuum of Macro-molecular Assembly Models with Graph Templates of Complexes

T. Dreyfus * and V. Doye † and F. Cazals ‡

Project-Team ABS

Research Report n° 8118 — October 2012 — 18 pages

Abstract: Reconstruction by data integration is an emerging trend to reconstruct large protein assemblies, but uncertainties on the input data yield average models whose quantitative interpretation is challenging. This paper presents methods to probe fuzzy models of large assemblies against atomic resolution models of sub-systems.

More precisely, consider a Toleranced Model (TOM) of a macro-molecular assembly, namely a continuum of nested shapes representing the assembly at multiple scales. Also consider a template namely an atomic resolution 3D model of a sub-system (a complex) of this assembly. We present graph-based algorithms performing a multi-scale assessment of the complexes of the TOM, by comparing the pairwise contacts which appear in the TOM against those of the template. We apply this machinery to recent average models of the Nuclear Pore Complex, and confront our observations to the latest experimental work.

The software implementing the algorithms of this paper, GRAPH_MATCHER, is available from the VORATOM suite, see http://team.inria.fr/abs/software/voratom.

Key-words: Macro-molecular assembles, Reconstruction by data integration, Nuclear Pore Complex, Model assessment, Toleranced Models, isomorphic graphs, maximum common sub-graphs

* INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; email: Tom.Dreyfus@sophia.inria.fr

† Institut Jacques Monod, CNRS, UMR 7592, Univ Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France; email: doye.valerie@ijm.univ-paris-diderot.fr

‡ INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; email: Frederic.Cazals@sophia.inria.fr

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Sur la cohérence d'un continuum de modèles d'un assemblage macro-moléculaire et de graphes de sous-complexes

Résumé : La reconstruction par intégration de données est une modalité émergente pour reconstruire de gros assemblages macro-moléculaires, mais les incertitudes sur les entrées donnent lieu à la génération de modèles moyens dont l'interprétation quantitative est délicate. Ce travail présente des méthodes pour comparer de tels modèles moyens à des structures de sous-systèmes connus à résolution atomique.

Plus précisément, considérons un modèle tolérancé (TOM) d'un assemblage, i.e. un continuum de formes imbriquées représentant l'assemblage à diverses échelles. Considérons également un *template*, i.e. un modèle à résolution atomique d'un sous-système. Nous présentons des outils dérivés de la théorie des graphes, permettant de comparer les contacts entre les protéines du TOM aux contacts du template. Nous utilisons ces outils pour analyser des modèles moyens du pore nucléaire récemment produits, et discutons nos résultats à la lumière des données expérimentales les plus récentes.

Le logiciel implémentant les algorithmes de ce travail, baptisé GRAPH_MATCHER, est disponible au sein de l'environnement logiciel VORATOM, voir http://team.inria.fr/abs/software/voratom.

Mots-clés : assemblage macro-moléculaire, reconstruction par intégration de données, pore nucléaire, évaluation de modèles, modèles tolérancés, graphes isomorphes, plus grands graphes communs

1 Introduction

Reconstruction by data integration. Small protein complexes involving a handful of polypeptide chains are typically studied with X ray crystallography and/or NMR. The situation is completely different for macro-molecular assemblies involving from tens to hundreds of chains (nuclear pore complex, chaperonin cavities, the proteasome, ATP synthases, etc), for which the impossibility to obtain thorough information using the aforementioned experimental techniques motivated the development of *reconstruction by data integration (RDI)* [AFK+08]. In RDI, plausible coarse-grain models of the assembly, which typically use of the order of 10 balls per protein (or nucleic acid), are selected by a non-convex optimization function maximizing the agreement of the model with various experimental data (cryo-electron microscopy (cryo-EM), centrifugation, immuno-labeling experiments, tandem affinity purification, etc).

However, both the data used in RDI and the results provided need to be handled with care. For example, the maps obtained in cryo-EM are often noisy, so that choosing a density level to contour a surface enclosing the model is not trivial [Fra06]. Protemomics data are often difficulty to interpret. In particular Tandem Affinity Purification Data (TAP) [PCR+01] , which provide proximity information between protein types, are inherently ambiguous since the protein types identified do not provide any information on the number of complexes involving the tagged protein—letting alone the stoichiometry issues. Immuno labeling experiments, which consist of locating a protein thanks to specific immunoglobulins tagged with gold particles, suffer from localization uncertainties. Other types of data present similar issues.

Consequently, the problem of selecting models which best comply with the data is inherently ill-posed, and the output typically consists of an ensemble of plausible models. This state of affairs is illustrated by the reconstruction of the Nuclear Pore Complex (NPC) [ADV+07a, ADV+07b], a protein assembly with eight-fold axial symmetry regulating the nucleo-cytoplasmic transport, and made of $\sim$ 456 protein instances of 30 protein types. Based on the reconstruction by data integration approach, the authors were able to select 1000 coarse-grain structures. These structures were further averaged to produce one *probability density map* per protein type, i.e. a map encoding the probability of presence of the instances of that type in the NPC.

Parameterized assembly models and their assessment. The maps produced for the NPC are of probabilistic nature and actually mirror the uncertainties on the input data [ADV+07a, ADV+07b]. For a given map, while identifying high confidence landmarks is rather easy, e.g. the local maxima of the probability, making statements of the rest of the map is more challenging. For example, the mere problem of placing a prescribed number of protein instances within a map is an ill-posed problem, as the volume of the union of the voxels with a strictly positive value can be significantly larger than the volume of these instances, a fact which motivated the introduction of Toleranced Models (TOM) [LWC97]. Prosaically, a TOM is a mathematical model whose modeling primitive is toleranced ball, namely two nested concentric balls [CD10]: the inner ball represents a high confidence region, while the outer one delimits the range of interest. The exploration of the region between these two balls is achieved by interpolating between the inner and the outer radii, a process governed by a parameter $\lambda \in [0.1]$. Because a TOM inherently defines a continuum of model, probing this continuum against an atomic resolution structure or a structure within which the pairwise contacts are believed to be correct is a key endeavor.

TOM: previous work and contribution. TOM were introduced in [CD10], and two types of analysis were developed in [DDC12]. We first introduced the notion of contact probability, namely the probability to observe k dimers involving two specific protein types, as a function of λ. Second, we analyzed *isolated copies* i.e. protein complexes involving a given set of protein

types with prescribed stoichiometry, and discussed their number with respect to the symmetry properties of the assembly—this latter analysis relies on so-called *Hasse diagrams*, see Fig. 1(D) in [CD10].

These analysis actually suffer from two main limitations. First, isolated copies do not allow reporting protein complexes which involve the given set of protein types but whose stoichiometry differs from the imposed one. Second, the connectivity of a protein complex cannot be assessed, so that two complexes having the same stoichiometry (for each protein species) cannot be distinguished even if they have different connectivity.

In the sequel, we provide tools circumventing these limitations, with one key application in mind, namely the comparison of a protein complex from a TOM model, against a high resolution model obtained by crystallography and/or modeling.

2 Methods

2.1 Hasse Diagram and Isolated Copies

Since our tools shall be applied to nodes of the Hasse diagram of a TOM, we first recall some fundamental notions (see also supplemental section in [DDC12]).

The growth process mentioned in introduction results in merges between growing proteins and complexes, which occur for specific values of the parameter λ (see Fig. 1(C) in [DDC12]). We record such events in a directed acyclic graph, also called a Hasse diagram: its nodes correspond to protein complexes; an edge corresponds to an *ancestor* - *successor* pair in the merge process. Prosaically, the Hasse diagram records the protein contact history (see Fig. 1(D) in [DDC12]). The Hasse diagram can be constructed in the *monocolor setting* where all the protein types are indistinguishable. But it can also be constructed in the *bicolor setting*, where some protein types of interest have been painted in red, the remaining types being painted in blue. For example, the red proteins may represent the types seen in a Tandem Affinity Experiment, or the types present in a complex. In this bicolor setting, given a stoichiometry for each red protein type, a node of the Hasse diagram is called *an isolated copy* if it contains the prescribed number of instances of each red protein type.

Note that the larger the value of λ at which a complex appears, the weaker the accuracy of the contacts between its constituting instances. To quantify this accuracy, the volume ratio $\overline{V}_\lambda(C)$ measures the ratio between the volume of the toleranced balls of the protein complex C interpolated at λ, and a reference volume of C computed from the sequences of the proteins [DDC12].

2.2 Comparing a Protein Complex to a Template

In the sequel, we assume that a Hasse diagram associated to a set of protein types in a TOM is given. Denote C the protein complex associated to a node of the Hasse diagram. We endow this node with its *skeleton graph* G_C, which encodes the pairwise contacts within C. We wish to compare the skeleton graph G_C against the skeleton graph G_t of a template T of C (Fig. 1). Practically, T shall be a co-crystallized complex or a high-resolution model built in-silico, and the protein types in T identify the red proteins of the bicolor setting.

Search of protein complexes similar to a template. The skeleton graph G_C corresponds to a complex C whose nodes are protein instances i.e. each instance carries a unique identifying label in the assembly. On the one hand, the nodes of G_t are protein types; assuming that the

template contains at most one instance of each type, a node of G_C (a protein instance) can be uniquely mapped to a node of G_t (a protein type).

We assume that all the types of the instances present in the protein complex C are present in the template skeleton graph T. But the complex C may not feature instances of all the types found in the template T. We therefore denote $G_{t|C}$ the *restricted template* i.e. the graph obtained by removing from G_t all the nodes whose protein types are not found in the protein instances of G_C, and the edges incident on these nodes. To compare the graphs $G_{t|C}$ and G_C, we use the concept of *matching*. Since matchings are intimately related to Maximal Common Induced Sub-graphs (MCIS) and Maximal Common Edge Sub-graphs (MCES), we first recall these notions (Supplemental Fig. 4):

Maximal Common Sub-graphs. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two undirected labelled graphs.

Definition. 1. *A* **Maximal Common Edge Sub-graph (MCES)** *of G_1 and G_2 is a graph H which is isomorphic to sub-graphs G'_1 of G_1 and G'_2 of G_2, such that there is no other Common Edge Sub-graph H' of G_1 and G_2 containing H.*

Definition. 2. *An* **induced sub-graph** *G' of G is a sub-graph of G such that for all pairs of vertices (u, v) of G', (u, v) is an edge of G' iff it is an edge of G.*

Definition. 3. *A* **Maximal Common Induced Sub-graph (MCIS)** *of G_1 and G_2 is a graph H which is isomorphic to* induced sub-graphs *G'_1 of G_1 and G'_2 of G_2, such that there is no other Common Induced Sub-graph H' of G_1 and G_2 containing H.*

Notice in particular that a MCES or MCIS calculation yields in general several matchings. Computationally, the computation of MCES and MCIS is tantamount to the computation of maximal cliques [CK05].

2.3 Matchings: Depleted, Complete, Exact and Perfect

Matchings and their signatures. We define a *matching* as the mapping between vertices and edges of $G_{t|C}$ and G_C, defined either by a MCIS or a MCES. That is, the matching maps vertices of $G_{t|C}$ (protein types of the template) to vertices of G_C (protein instances of the complex), and edges of $G_{t|C}$ (contacts within the template) to edges of G_C (contacts within the complex). Taking the template as reference, we assess a matching with its signature (Fig. 1):

- *Matched protein type(s):* a protein type of $G_{t|C}$ with a corresponding instance in G_C. This set is denoted $V^{\sim}$.
- *Missing protein type(s):* a protein type of G_t with no corresponding instance in G_C. This set is denoted V^{-}.
- *Matched contact(s):* a contact in $G_{t|C}$ with a counterpart in G_C. This set is denoted $E^{\sim}$.
- *Missing contact(s):* a contact in $G_{t|C}$ whose protein types match instances in G_C but with no corresponding contact in G_C. This set is denoted E^{-}.
- *Extra contact(s):* a contact in G_C whose protein instances match types in $G_{t|C}$ but with no corresponding contact in $G_{t|C}$. This set is denoted E^{+}.

Using these sets, the *signature* of the matching A is defined by:

$$S(G_t; G_C; A) = \{V^{\sim}, V^{-}, E^{\sim}, E^{-}, E^{+}\}. \tag{1}$$

To further classify matchings, we use the two sets V^- and $E^- \cup E^+$. Each such set faces two situations, namely is empty or not, which yields four types of matchings:

Definition. 4. *A matching is called* depleted *when*

$$V^- \neq \emptyset \text{ and } E^- \cup E^+ \neq \emptyset \tag{2}$$

A matching is called complete *when*

$$V^- = \emptyset \text{ and } E^- \cup E^+ \neq \emptyset \tag{3}$$

A matching is called exact *when*

$$V^- \neq \emptyset \text{ and } E^- \cup E^+ = \emptyset \tag{4}$$

A matching is called perfect *when*

$$V^- = \emptyset \text{ and } E^- \cup E^+ = \emptyset \tag{5}$$

Note that for an exact matching, all the contacts between the protein types of $V^{\sim}$ are present, but some protein types are missing, i.e. contacts incident to missing protein types are not considered as missing contacts.

2.4 Assessing a Template in a Hasse Diagram

Depleted and Complete matchings. To maximize the number of common contacts between the skeletons of complexes C of the Hasse diagram and a given restricted template $G_{t|C}$, we carry out MCES calculation as follows. First, for each complex C which is as root of the Hasse diagram, we compute the MCES between G_C and $G_{t|C}$. Second, let A be a matching returned by the MCES calculation. We search in the Hasse diagram the ancestor D of C involving the protein instances and contacts of C matched by A, and which minimizes the number of extra contacts.

Exact and Perfect matchings. An exact matching associated to a node C of a Hasse diagram is maximal provided that there does not exist in the Hasse diagram any exact matching for a successor of C. Such complexes are easily obtained from the matchings provided by a MCIS calculation between the graphs G_C and $G_{t|C}$ in each node of the Hasse diagram.

The support and the instantiation of a matching. As just explained, matchings are sought by performing MCES and MCIS calculations. Consider a complex C for which a matching has been found. The complex associated to the node C is said to *host* the matching. Equivalently the complex is called the *support* of the matching, and abusing terminology, the node associated to C may also be called the support of the matching. If the protein instances involved in the matching form a sub-complex of C, these instances are called the *instantiation* of the matching. Note in particular that a given complex C may be the support of several matchings.

3 Material: the Nuclear Pore Complex

As system under scrutiny, we use the NPC, the largest protein assembly of the eukaryotic cell known to date, which consists of about $n \sim 456$ instances of $k = 30$ protein types, the stoichiometry of a protein type being 8 or 16. So far, two types of results have been obtained on the NPC. On the one hand, global yet qualitative results have been obtained based on data integration, in particular the aforementioned maps [ADV$^+$07a, ADV$^+$07b]. On the other hand, a number of complexes of the NPC have been modeled at atomic resolution, using in particular crystal structures of dimers or trimers. This is in particular the case of the so-called Y-complex, an heptameric complex playing a major role to form the NPC scaffold [LKB$^+$02] (Fig. 2(A)). Yet, the embedding of its 16 [ADV$^+$07b] (or possibly 32 [DMS$^+$08]) copies within the NPC remains under controversy with three competing models: the first one supports the presence of two closed rings containing each eight copies of the Y-complex arranged in a head-to-tail manner [ADV$^+$07b, SMD$^+$09, M. 11]; in the second one (lattice model), the two sets of eights copies display a vertical head-to-head orientation [BLS$^+$08, BS12]; finally a fence-like arrangement based on the organization of Seh1-Nup85 as hetero-octamer was proposed [DMS$^+$08]. While this question is still under debate, the interfaces i.e. the pairwise contacts reported in [FMPS$^+$12] are incompatible with the third model.

Speaking of these pairwise contacts, six of them involving the seven protein species were previously established (in the sequel, we use the nomenclature defined in [BLS$^+$08]); in particular, two dimers (the Y_X-long-arm involving (Nup85,Seh1)[BLS$^+$08], and the Y_X-tail involving (Nup133,Nup84)[WS09]), and one trimer (the Y_X-edge involving (Nup84,Nup145C, Sec13) [NHD$^+$09]) were solved by crystallography. These contacts define the template skeleton $G_t(Y)$ (Fig. 2(A)). Moreover, an additional contact between Nup145C and Nup85 was recently proposed in [FMPS$^+$12] and further characterized in [BS12]. Since this last contact was never observed in the TOM of the NPC (see Fig.4 in [DDC12]), we define the template skeleton $G_t(Y)$ from the six previous pairwise contacts (Fig. 2(A)).

Since our previous analysis [DDC12] focused on the Y-complex as a whole, we now use our graphical tools to assess the coherence between the TOM of the whole NPC and the pairwise contacts just recalled.

4 Results: Y-complex Analysis

4.1 Rationale

The analysis presented in [DDC12], based on the tools recalled in section 2.1, concluded that 11 isolated copies of the Y-complex were present in the TOM, out of the 16 expected (Fig. 2(B, C)). To understand this discrepancy, recall that the Y-complex consists of a tail known as the edge element E_i = (Y_X-tail,Y_X-edge) and two arms A_i = (Nup120, Y_X-long-arm) (Fig. 2(A)). Also recall that in the ring model, 8 copies of the Y-complex form an annulus on the cytoplasmic (and likewise on the nuclear side), so that these 8 copies can be cyclically ordered.

Now, let us focus on one of these 8 complexes in the TOM, say the ith one. If, along the growth process of the TOM, the merge of the arms A_{i-1} with the edge E_i (, or that of the arms A_i with the edge E_{i+1},) occurs before the merge between E_i and A_i, then, this ith copy does not appear as an isolated copy since the stoichiometry condition is violated.

We now show how matchings circumvent this limitation, and also provide information of pairwise contacts. Along the way, we shall refer to the matchings of the Table 3(A,B,C), each of them being referred to by a tag, e.g. $M_D(i)$ for a depleted matching, $M_C(i)$ for a complete matching, $M_E(i)$ for an exact matching, and $M_P(i)$ for a perfect matching.

4.2 Depleted and Complete Matching

To identify the 16 copies of the Y-complex, we first compute the depleted and complete matchings of the protein complexes at the roots of the Hasse diagram with $G_{t|C}(Y)$ (Fig. 2 (C, (a) and (b), and Fig. 3(A,B)). Ten depleted matchings are obtained for (Y_X-tail,Y_X-edge) ($M_D(1), M_D(2)$), and likewise for (Nup120 ,Y_X-long-arm) ($M_D(3)$). These matchings correspond to ten instances of the Y-complex split into two complexes—that is another protein instance (from this precise Y-complex, or from another Y-complex, or another NUP) disrupts the connectivity (Nup145C, Nup120). To value the interest of matchings, we also counted the number of matchings having a support which is one of the 11 isolated copies of the Y-complex (i.e, the green nodes in Fig. 2(C)). Counts of 0 and 1 ($M_C(1)$) for the depleted and complete matchings illustrate the radically different nature of the information encoded by both constructions.

By inspecting the instantiations of the depleting matchings $M_D(1), M_D(2)$ and $M_D(3)$, it turns out that each of the ten split Y-complex admits two possible reconstructions. Indeed, as explained in Sec. 4.1, one does not know whether two successive pieces are part of the same Y-complex instance or of two distinct instances. Further inspection of the six matchings for the whole Y-complex (two depleted ($M_D(4)$ with only one missing protein type, and four complete matchings $M_C(1)$)) shows that once embedded into the two rings (Fig 1(B)), five Y-complex instances have the same orientation in a ring, the last one having the opposite orientation in the other ring. Note that the split Y-complex instances neighboring a complete Y-complex instance have only one possible reconstruction. Phrased differently, the complete Y-complexes impose a unique configuration for each of the 16 instances of the Y-complex.

Inspection of the signatures of all these matchings also shows that the number of extra contacts observed is bounded by seven for a maximum of fifteen, see $\max E^+$ column in Fig. 3(B). Indead, seven proteins make at most twenty one pairwise contacts, out of which six belong to the template.

To assess the geometric accuracy of these complexes, we observe that the volume ratio does not exceed 5.83 ($M_D(3)$), namely twice the maximal volume ratio of the 11 isolated copies ($\overline{V}_\lambda \in [0.86, 2.14]$). This degradation in moving from isolated copies to depleted or complete matchings is actually expected: the connectivity constraint specifying a matching being more stringent than the mere presence constraint qualifying an isolated copy, larger values of λ (whence larger volumes) are required to meet the constraint imposed by the template skeleton.

4.3 Exact and Perfect Matching

Since we did not observe any perfect matching, we successively analyze the largest exact matchings obtained, and proceed with those corresponding to crystallographic complexes. (Fig. 2 (C, (c)), and Fig. 3(C))

Largest exact matchings. Five exact matchings involving four protein types ($M_E(1)$ with Nup84, Nup145C, Nup120 and Nup85) are found, but no exact matching containing a superset of these proteins is observed.

We also note that a single matching in $M_E(1)$ has as support which is an isolated copy, this copy being also the support of the complete matching $M_C(1)$ (bicolored node in Fig. 1(C, (b) and (c))).

Y_X-tail. We found 16 exact matchings containing the Y_X-tail ($M_E(1)$+$M_E(2)$), 15 of them strictly matching the Y_X-tail ($M_E(2)$), and a supplementary one also containing Nup145C ($M_E(3)$). The absence of exact matching containing the Y_X-tail (if one omits $M_E(3)$) shows that instances of Nup145C make spurious contacts with Nup133 before Nup84 during the growth process. We note in passing that these extra contacts prevent from finding exact matchings but

do not provide any hindrance for depleted and complete matchings—this actually owes to the difference between MCIS and MCES.

We also observe that no exact matching with the Y_X-tail whose support is an isolated copy.

Y_X-edge. We found only five exact matchings containing the Y_X-edge ($M_E(4)$). However, we found 18 exact matchings containing the contact (Nup145C, Nup84) (five in $M_E(1)$, one in $M_E(3)$, five in $M_E(4)$ and seven in $M_E(5)$). Since our previous study based on contact probabilities revealed that the 16 instances of Nup84 are connected to instances of Nup145C (see Fig. 4(A) and (C) in [DDC12]), selected instances of Nup84 or Nup145C actually appear in several exact matchings, reflecting their poor positioning. Also, there are six exact matchings containing exactly Nup145C and Sec13 ($M_E(6)$): since we observe the 16 instances of the contacts (Nup84, Nup145C), we deduce that these six instances of Sec13 and six instances of Nup84 make an extra contact. Four instances of Sec13 making no valid contact with Nup145C are evidenced ($M_E(8)$): while the poor placement of instances of Sec13 had previously been established using contact probabilities [DDC12], the matchings $M_E(8)$ provide additional information on the missed contacts.

Finally, we found one exact matching with four protein types (Sec13, Nup145C, Nup120, Nup85) ($M_E(7)$) whose support is an isolated copy, namely that also supporting the complete matching $M_E(1)$).(bicolored node in Fig. 1(C, (b) and (c)))

Y_X-long-arm. There are 16 exact matchings containing exactly the Y_X-long-arm ($M_E(9)$): since our previous analysis revealed that there are 16 instances of the contact (Nup85, Nup120) (see Fig. 4 in [DDC12]), an extra contact exists between the 16 instances of Seh1 and Nup120. We also note that there are two exact matchings ($M_E(9)$) having as support two different isolated copies.

Nup120. We found 17 exact matchings with Nup120 (five in $M_E(1)$, one in $M_E(5)$, ten in $M_E(10)$ and one in $M_E(11)$). In other words, distinct instances of Nup120 are present in several exact matchings with different restrained templates.

5 Discussion

The notions of isolated copy and contact frequency introduced in [DDC12] improved the contact frequencies and subsequent analysis presented in [ADV+07a, ADV+07b], allowing in particular to make a quantitative assessment of pairwise contacts and protein complexes in large protein assembly reconstructed by data integration. Yet, connectivity analysis were beyond reach, a limitation circumvented by the graph-based tools presented in this paper.

Regarding the Y-complex, we were able to reconstruct the 16 existing copies of the Y-complex in the TOM of the NPC using depleted and complete matchings. However, at least half of the instances of the Y-complex are split in two pieces due to the disruption of the contact (Nup145C, Nup120). We also show using the exact matchings that all protein contacts observed in crystal structures of complexes of the Y-complex are present in our toleranced model. However, with the exception of an exact matching involving Y_X-tail ($M_E(3)$), no exact matching strictly contains the template of a crystallographic dimer or trimer. Instead, larger complexes are found as pairs of smaller exact matchings. In fact, the five exact matchings $M_E(1)$ show that with a better positioning of Sec13 and Seh1, a exact matching involving the union of crystal based templates Nup120, Y_X-long-arm and Y_X-edge would be obtained.

From the modeling standpoint, the methods presented in this paper complement those developed in our initial work [DDC12], based on isolated copies. While isolated copies focus on the separability and the lifetime of complexes involving prescribed protein types, matchings aim at

assessing the connectivity of protein instances involving the contacts of a template. Ideally, an isolated copy should support matchings, a failure of this property witnessing a situation where the connectivity of the copy differs radically from that of the template skeleton. Reciprocally, a matching whose instantiation does not belong to an isolated copy corresponds to a complex which is not well separated from the protein types defining the isolated copies.

All in all, our tools serve two purposes in the context of reconstruction by data integration. On the one hand, selecting the models which best comply with experimental data is important in a complex modeling pipeline, as such decisions are likely to improve the convergence of the optimization process. On the other hand, the ability to test any hypothesis, in terms of protein contacts, may orientate experiments so as to confirm or falsify putative 3D models. That is, we believe that our tools are particularly well suited to leverage reconstruction by data integration, by initiating a virtuous loop mixing modeling and experiments.

References

[ADV+07a] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[ADV+07b] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, A. Sali, and M.P. Rout. The Molecular Architecture of the Nuclear Pore Complex. *Nature*, 450(7170):695–701, Nov 2007.

[AFK+08] F. Alber, F. FÃ¶rster, D. Korkin, M. Topf, and A. Sali. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[BLS+08] S.G. Brohawn, N.C. Leksa, E.D. Spear, K.R. Rajashankar, and T.U. Schwartz. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science*, 322:1369–1373, 2008.

[BS12] S. Bilokapic and T. U. Schwartz. Molecular basis for nup37 and ely5/elys recruitment to the nuclear pore complex. *PNAS*, 109(38):15241–15246, 2012.

[CD10] F. Cazals and T. Dreyfus. Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, Lyon, 2010.

[CK05] F. Cazals and C. Karande. An algorithm for reporting maximal c-cliques. *Theoretical Computer Science*, 349(3):484–490, 2005.

[DDC12] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macro-molecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 2012. In press.

[DMS+08] E.W. Debler, Y. Ma, H.-S. Seo, K.-C. Hsia, T.R. Noriega, G. Blobel, and A. Hoelz. A fence-like coat for the nuclear pore membrane. *Molecular Cell*, 32:815–826, 2008.

[FMPS+12] J. Fernandez-Martinez, J. Phillips, M.D. Sekedat, R. Diaz-Avalos, J. Velazquez-Muriel, J.D. Franke, R. Williams, D.L. Stokes, B.T. Chait, A. Sali, and M.P. Rout. Structure–function mapping of a heptameric module in the nuclear pore complex. *The Journal of Cell Biology*, 2012.

[Fra06] J. Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state.* Oxford University Press, USA, 2006.

[LKB+02] M. Lutzmann, R. Kunze, A. Buerer, U. Aebi, and E. Hurt. Modular self-assembly of a y-shaped multiprotein complex from seven nucleoporins. *The EMBO Journal*, 21(3):387–397, 2002.

[LWC97] J-C. Latombe, R. Wilson, and F. Cazals. Assembly sequencing with toleranced parts. *Computer Aided Design*, 29(2):159–174, 1997.

[M. 11] M. Kampmann and C.E. Atkinson and A.L. Mattheyses and S.M. Simon. Mapping the Orientation of Nuclear Pore Proteins in Living Cells with Polarized Fluorescence Microscopy. *Nat. Struct. Mol. Biol.*, 18(6):643–652, 2011.

[NHD+09] V. Nagy, K.-C. Hsia, E.W. Debler, M. Kampmann, A.M. Davenport, G. Blobel, and A. Hoelz. Structure of a trimeric nucleoporin complex reveals alternate oligomerization states. *PNAS*, 106(42):17693, 2009.

[PCR+01] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. The tandem affinity purification method: A general procedure of protein complex purification. *Methods*, 24:218–229, 2001.

[SMD+09] H.-S. Seo, Y. Ma, E.W. Debler, D. Wacker, S. Kutik, G. Blobel, and A. Hoelz. Structural and Functional Analysis of Nup120 Suggests Ring Formation of the Nup84 Complex. *PNAS*, 106(34):14281–14286, 2009.

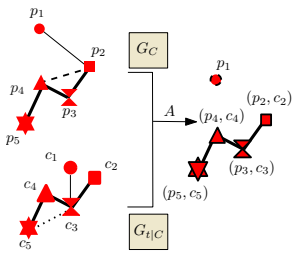
[WS09] J.R.R. Whittle and T.U. Schwartz. Architectural nucleoporins Nup157/170 and Nup133 are structurally related and descend from a second ancestral element. *J. Biol. Chem.*, 284:28442–28452, 2009.

Figure 1 Signature of a matching between the skeleton graph G_C of a complex C and a restricted template $G_{t|C}$. The match between a protein instance of G_C and a type of $G_{t|C}$ is materialized by an identical geometric shape (disk, square, triangle, hourglass). For nodes, bold contours indicate matching protein types, while dashed contours indicate missing protein types. For edges, bold lines indicate matching contacts, dashed lines indicate missing contacts, and dotted lines indicate extra contacts. Note that the adjectives matching/missing/extra qualify G_C w.r.t. $G_{t|C}$. **(A)** Depleted matching: a Maximal Common Edge Sub-graph calculation yields a matching with at least one missing protein type, and missing contacts (dotted lines) and/or extra contacts (dashed lines). **(B)** Complete matching: a Maximal Common Edge Sub-graph calculation yields a matching with no missing protein type. **(C)** Exact matching: a Maximal Common Induced Sub-graph calculation yields one or more matchings without any missing or extra edge.

(A)

Depleted matching:

$V^- \neq \emptyset$ and ($E^- \neq \emptyset$ or $E^+ \neq \emptyset$)

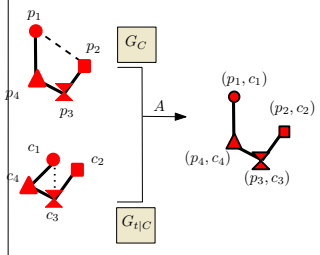


Signature \ Matching	A
$V^\sim$ Matching Protein Types	(p_2, c_2) (p_3, c_3) (p_4, c_4) (p_5, c_5)
V^- Missing Protein Types	c_1
$E^\sim$ Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_3, c_4) \leftrightarrow (p_3, p_4)$ $(c_4, c_5) \leftrightarrow (p_4, p_5)$
E^- Missing Contacts	(p_2, p_4)
E^+ Extra Contacts	(c_3, c_5)

(B)

Complete matching:

$V^- = \emptyset$ and ($E^- \neq \emptyset$ or $E^+ \neq \emptyset$)

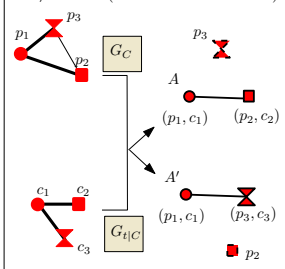


Signature \ Matching	A
$V^\sim$ Matching Protein Types	(p_1, c_1) (p_2, c_2) (p_3, c_3) (p_4, c_4)
V^- Missing Protein Types	$\emptyset$
$E^\sim$ Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_3, c_4) \leftrightarrow (p_3, p_4)$ $(c_1, c_4) \leftrightarrow (p_1, p_4)$
E^- Missing Contacts	(c_1, c_3)
E^+ Extra Contacts	(p_1, p_2)

(C)

Exact matching:

$V^- \neq \emptyset$ and ($E^- = \emptyset$ and $E^+ = \emptyset$)



Signature \ Matchings	A	A'
$V^\sim$ Matching Protein Types	(p_1, c_1) (p_2, c_2)	(p_1, c_1) (p_3, c_3)
V^- Missing Protein Types	c_3	c_2
$E^\sim$ Matching Contacts	(c_1, c_2) (p_1, p_2)	(c_1, c_3) (p_1, p_3)
E^- Missing Contacts	$\emptyset$	$\emptyset$
E^+ Extra Contacts	$\emptyset$	$\emptyset$

Figure 2 Analysis of the Y-complex in the TOM of the NPC: Hasse diagram, isolated copies and matchings. **(A)** Pairwise contacts between protein instances of the Y-complex, adapted from[BLS+08]; the template skeleton graph $G_t(Y)$ consists of nodes (colored blobs), connected by edges (black lines). The solved crystal structures of four Y-complex sub-complexes are also shown: the Y_X-short-arm [BLS+08], Nup120 [SMD+09], the Y_X-edge [NHD+09] and the Y_X-tail [WS09]. **(B)** TOM of the NPC at $\lambda = 0$ and $\lambda = 1$ restricted to the seven protein types of the Y-complex. An isolated copy of the Y-complex, involving one instance of each type, is colored on both pictures. Note that when λ increases, protein complexes (the grey domains) merge. **(C)** Hasse diagram of the Y-complex in the whole NPC, the three columns respectively presenting depleted, complete and exact matchings. The first row displays all the nodes hosting matchings of a given type (cyan for depleted, purple for complete, and dark blue for exact) accompanied by the number of associated matchings, except for the exact matchings where only nodes with multiplicity at least two are decorated to avoid cluttering. An isolated copy of the Y-complex is represented by a green node, except when it hosts a matching, in which case it is also circled by the color associated to the matching (purple for complete and blue for exact). The orange nodes correspond to the supports of the matchings singled out by the dashed arrow on the second row, each support being represented by colored balls. The edges representing contacts in the support are depicted with the conventions of Fig. 1 The last row presents the types and the contacts of the template with the same conventions. As discussed in section 2.4, note that the same node of the Hasse diagram may host several matchings — e.g. a depleted one (C,a) and an exact one (C,c).

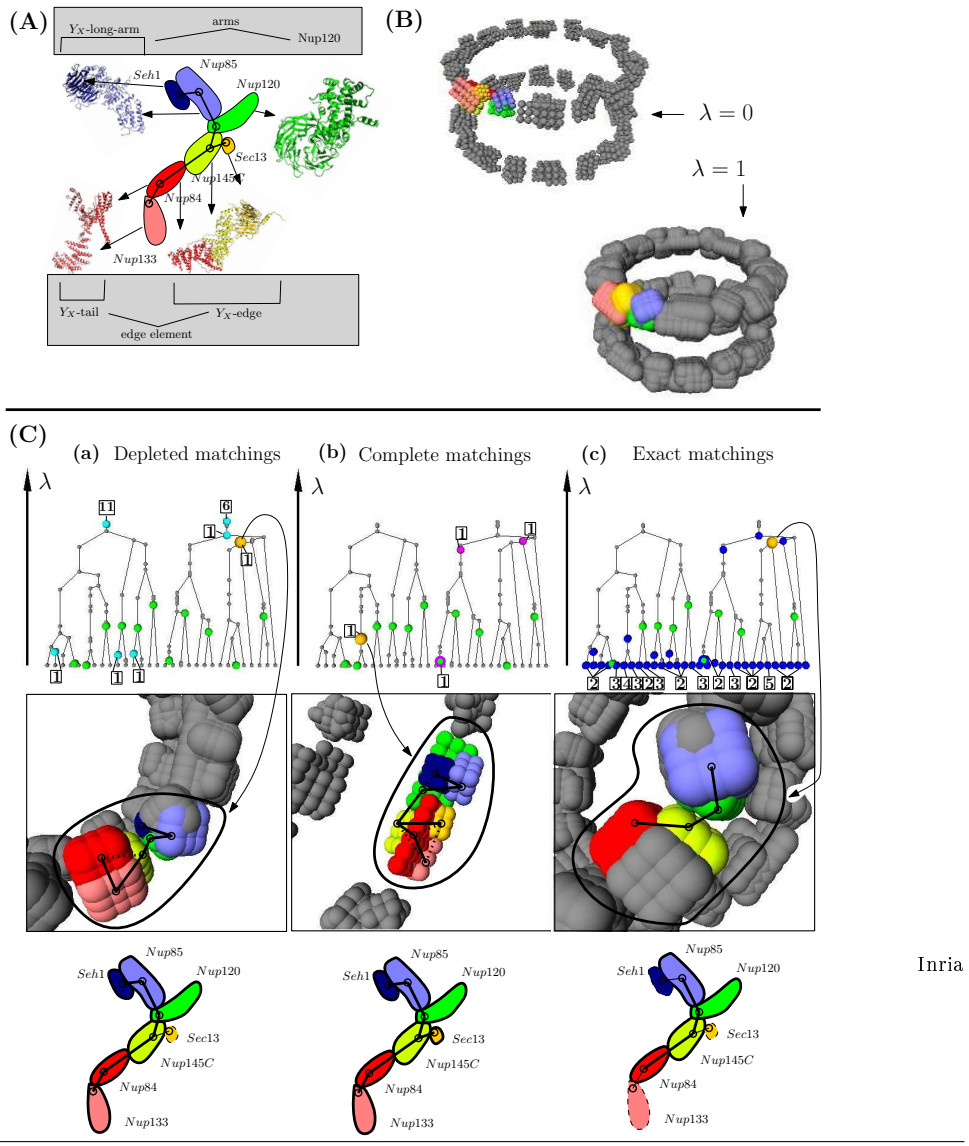
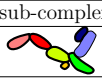


Figure 3 Using matchings to investigate the pairwise contacts within the instances of the Y-complex of the TOM. (A) All depleted matchings of protein complexes in the Hasse diagram with $G_t(Y)$. The table is split into three sections, delimited by double vertical bars. The first section gives the template skeleton, the tag of the depleted matching used in the text, and the number of depleted matchings with the same set of protein types. The second section provides the sizes of the sets involved in the signature of the matchings — for each matching we also indicate the number of possible extra contacts that is the number of pairs of matching protein types minus the number of matching contacts. The third section supplies the min and max volume ratios computed over the protein complexes associated to the concerned matchings—the volume ratio of a given complex being computed at the λ value at which this complex appears. **(B)** All complete matchings of protein complexes in the Hasse diagram with $G_t(Y)$. **(C)** All exact matchings of protein complexes in the Hasse diagram with $G_t(Y)$.

(A) Depleted matchings

Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid V^{-} \mid$	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	min $\mid E^{+} \mid$	max $\mid E^{+} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y)$;$M_D(1)$	2		4	2	0	1/1	1/1	4.49	4.71
$G_t(Y)$;$M_D(2)$	8		3	3	0	3/3	3/3	4.49	5.65
$G_t(Y)$;$M_D(3)$	10		4	2	0	1/1	1/1	1.17	5.83
$G_t(Y)$;$M_D(4)$	2		1	5	0	2/10	2/10	3.68	4.56

(B) Complete matchings

Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	min $\mid E^{+} \mid$	max $\mid E^{+} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y)$;$M_C(1)$	4		6	0	4/15	7/15	1.00	4.95

(C) Exact matchings

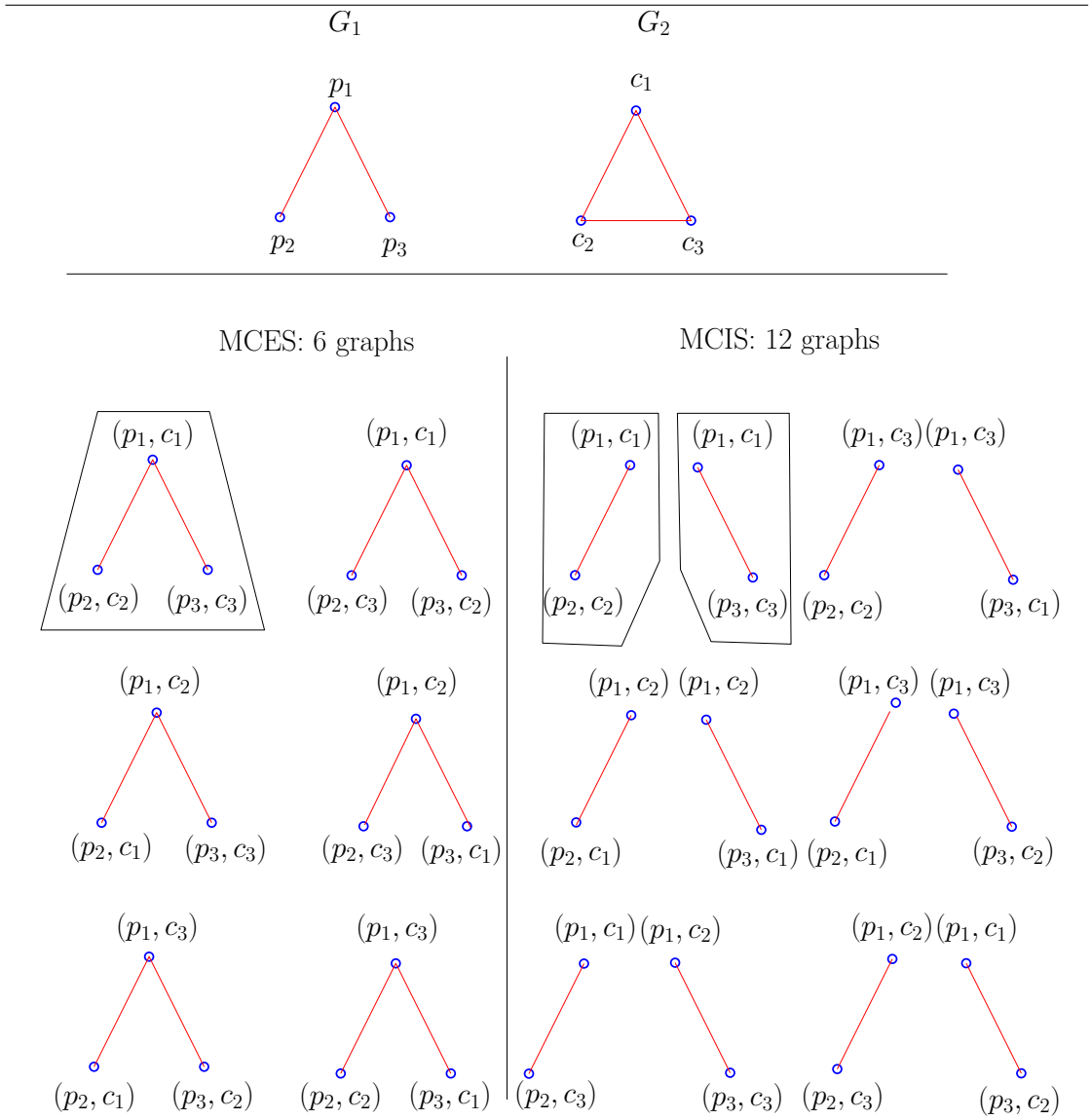
Template; tag	#	$V^{\sim}$ (sub-complexes)	$\mid V^{-} \mid$	min $\overline{V}_{\lambda}$	max $\overline{V}_{\lambda}$
$G_t(Y)$;$M_E(1)$	5		3	1.02	3.42
$G_t(Y)$;$M_E(2)$	15		5	0.77	0.89
$G_t(Y)$;$M_E(3)$	1		4	0.86	0.86
$G_t(Y)$;$M_E(4)$	5		4	0.78	0.86
$G_t(Y)$;$M_E(5)$	7		5	0.81	0.86
$G_t(Y)$;$M_E(6)$	6		5	0.80	0.84
$G_t(Y)$;$M_E(7)$	1		3	1.05	1.05
$G_t(Y)$;$M_E(8)$	4		6	0.54	0.63
$G_t(Y)$;$M_E(9)$	16		5	0.77	1.27
$G_t(Y)$;$M_E(10)$	10		5	0.88	0.91
$G_t(Y)$;$M_E(11)$	1		4	2.15	2.15

6 Supplement

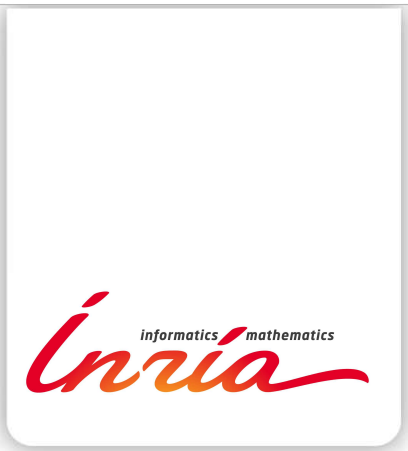
6.1 Graph theory

Computing matchings of two graphs G_1 and G_2 is tantamount to computing maximal cliques [CK05], and of particular interests are the matchings associated to the so-called Maximal Common Induced Sub-graph (MCIS) and Maximal Common Edge Sub-graph (MCES) of G_1 and G_2. These notions are illustrated on the Supplemental Figure 4. Notice in particular that a MCES or MCIS calculation yields in general several matchings.

Figure 4 Comparing graphs with matchings: illustration of the Maximal Common Edge Sub-graph (MCES) and Maximal Common Induced Sub-graph (MCIS) constructions. **Top.** Two labelled graphs G_1 and G_2. **Bottom Left.** The 6 MCES of G_1 and G_2 **Bottom Right.**: The 12 MCIS of G_1 and G_2. If we impose a correspondence, e.g. $((p_1 \leftrightarrow c_1), (p_2 \leftrightarrow c_2), (p_3 \leftrightarrow c_3))$ between the labels of the two graphs, there is one MCES and there are two MCIS — the circled graphs. Practically, the graph G_1 shall represent the skeleton graph G_C of a complex associated to a node of the Hasse diagram of a toleranced model, while the graph G_2 shall represent the skeleton graph $G_{t|C}$ of a template. In this context, the aforementioned correspondence reads as follows: the label p_i of G_C is identical to the label c_i of $G_{t|C}$, that is, they represent the same protein type.



Contents



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399